

# Dados Desbalanceados

Prof. André Gustavo Hochuli

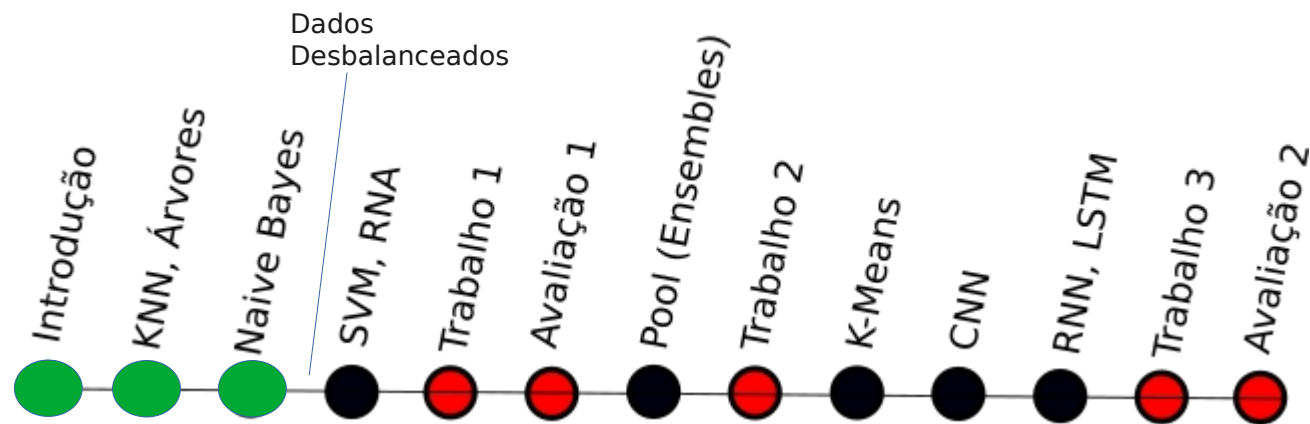
[gustavo.hochuli@pucpr.br](mailto:gustavo.hochuli@pucpr.br)

[aghochuli@ppgia.pucpr.br](mailto:aghochuli@ppgia.pucpr.br)

[github.com/andrehochuli/teaching](https://github.com/andrehochuli/teaching)

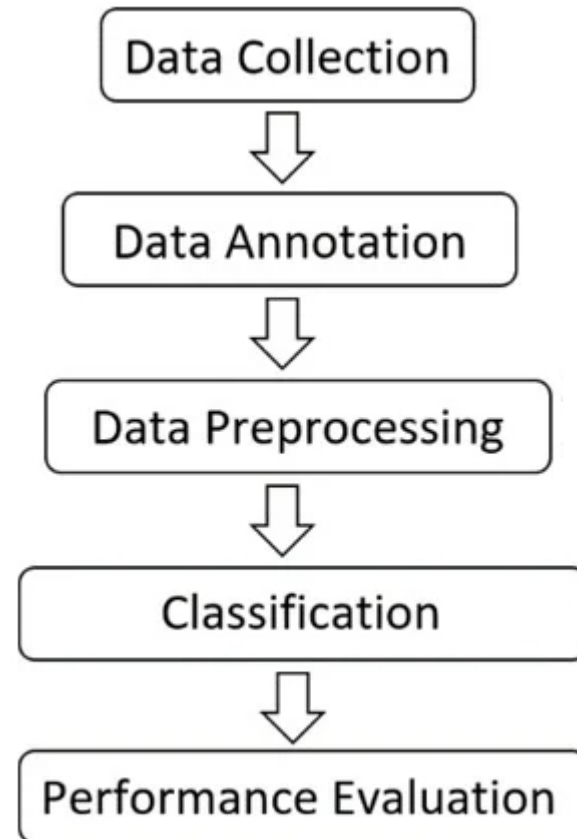
# Plano de Aula

- Discussões Iniciais
- Dados Desbalanceados
- Overfitting
- Exercícios



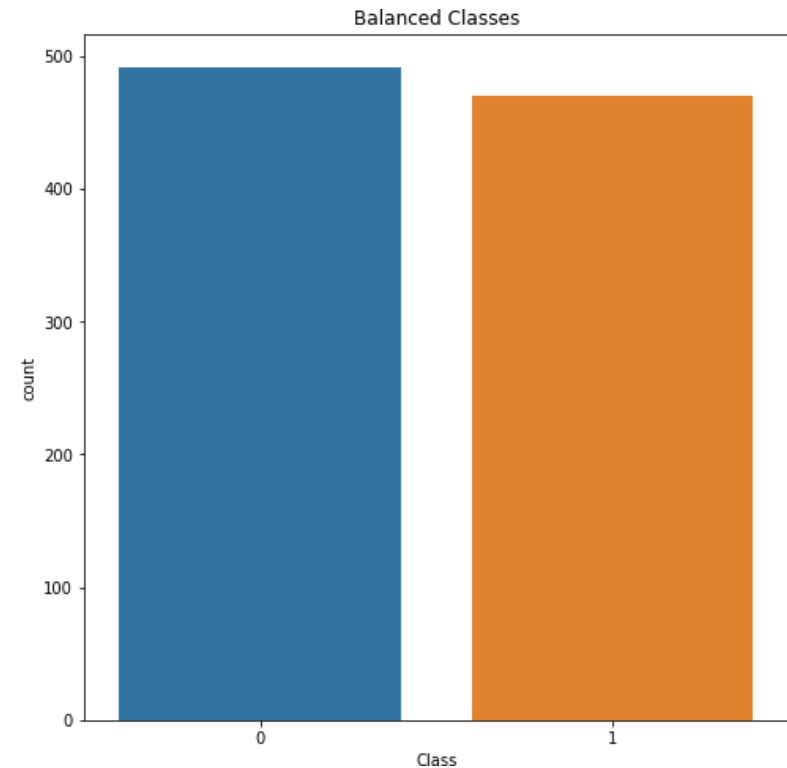
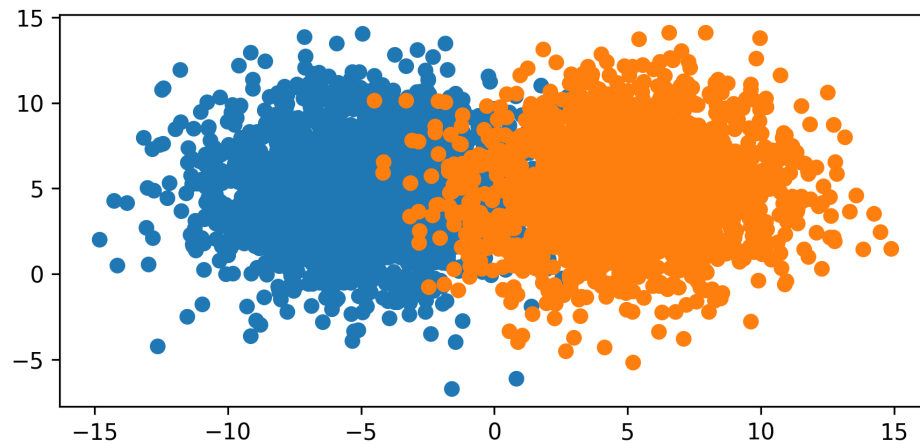
# Discussões Iniciais

- Modelos
  - KNN, NB e Árvores
- Anotação de Dados
  - Tarefa Manual
  - Representatividade
  - Dados Balanceados



# Dados Desbalanceados

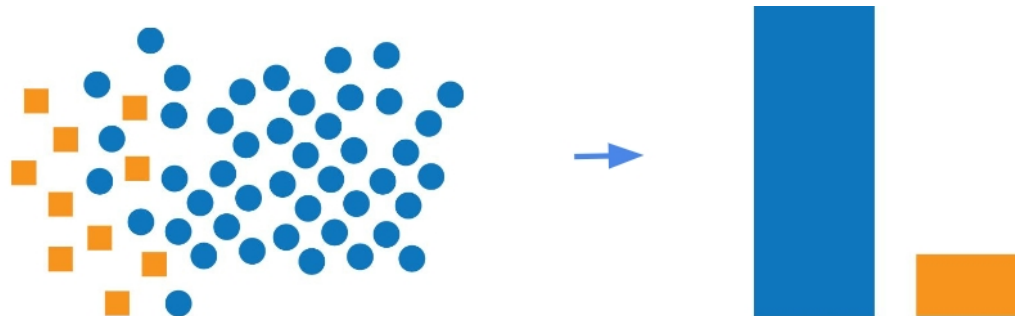
- Até o momento trabalhamos com datasets balanceados



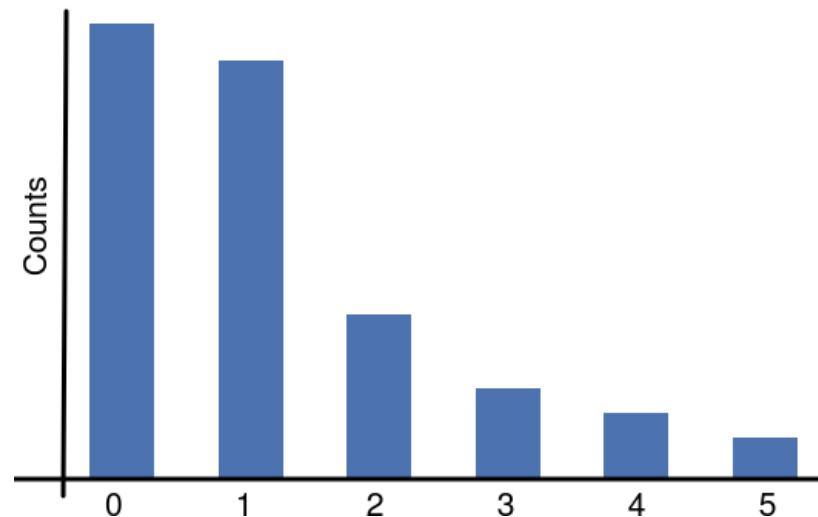
# Dados Desbalanceados

- No entanto, o mundo real nem sempre é balanceado

- Detecção de Fraudes
- Diagnósticos médicos



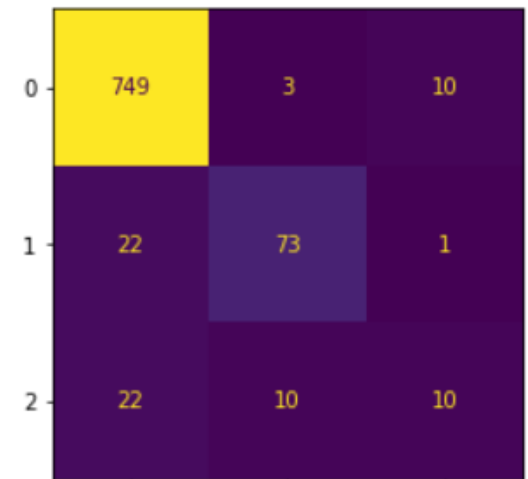
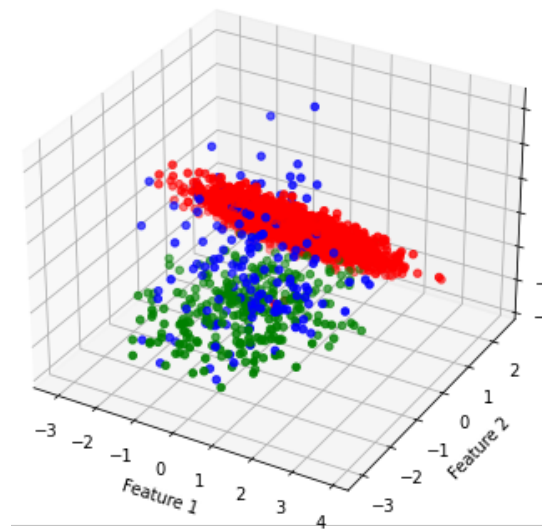
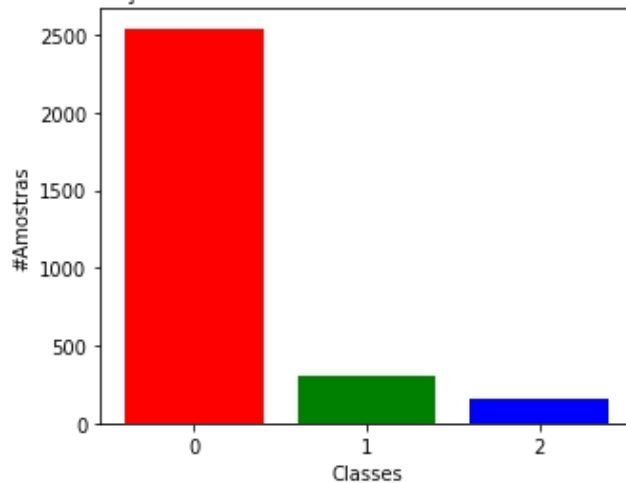
- Classificação de Espécies
  - 0 - Cachorro
  - 1 - Gato
  - ....
  - 5 - Mico Leão



# Dados Desbalanceados

- E qual o problema disso em machine learning?
  - Treinamento enviesado (generalização baixa)

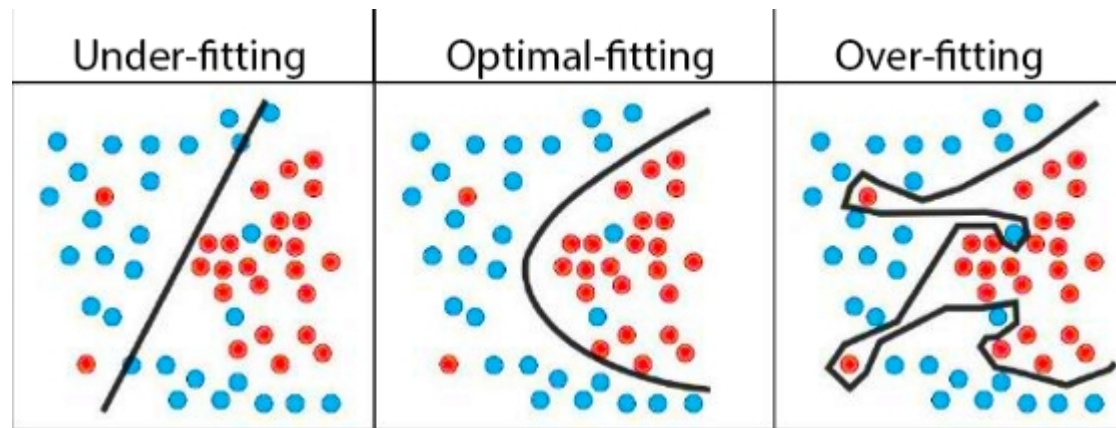
Distribuição das Classes - Sintetico - 3 Features - 3 Classes



	precision	recall	f1-score	support
0	0.94	0.98	0.96	762
1	0.85	0.76	0.80	96
2	0.48	0.24	0.32	42

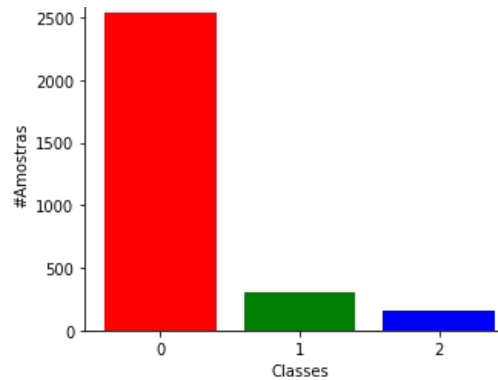
# Dados Desbalanceados

- E qual o problema disso em machine learning?
  - Overfitting

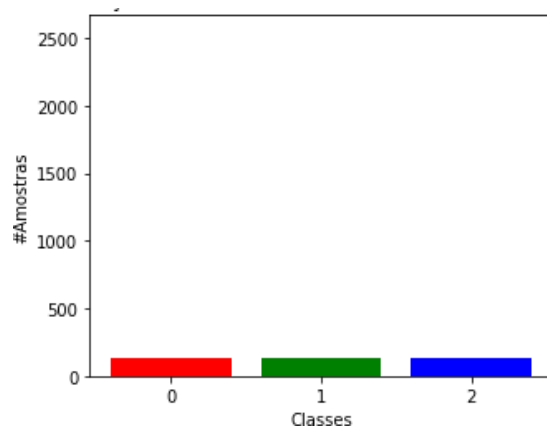


# Dados Desbalanceados

- E como tratar ?

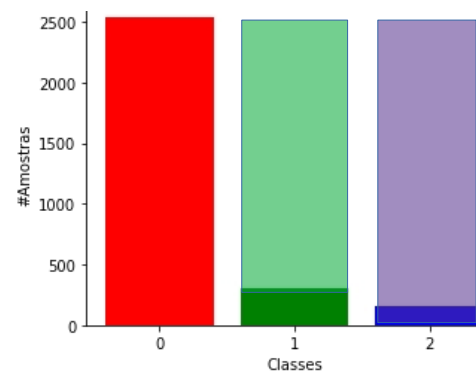


- Redução de Amostras (Undersampling)

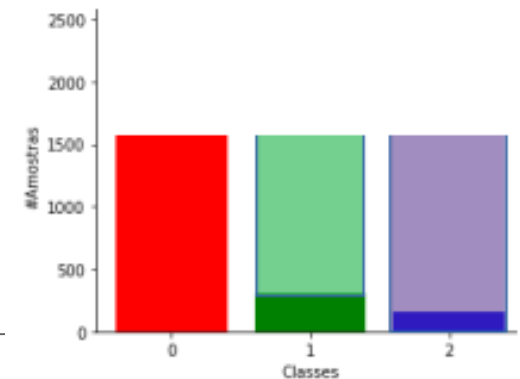


- Aumento de Amostras (Oversampling)

Replicação



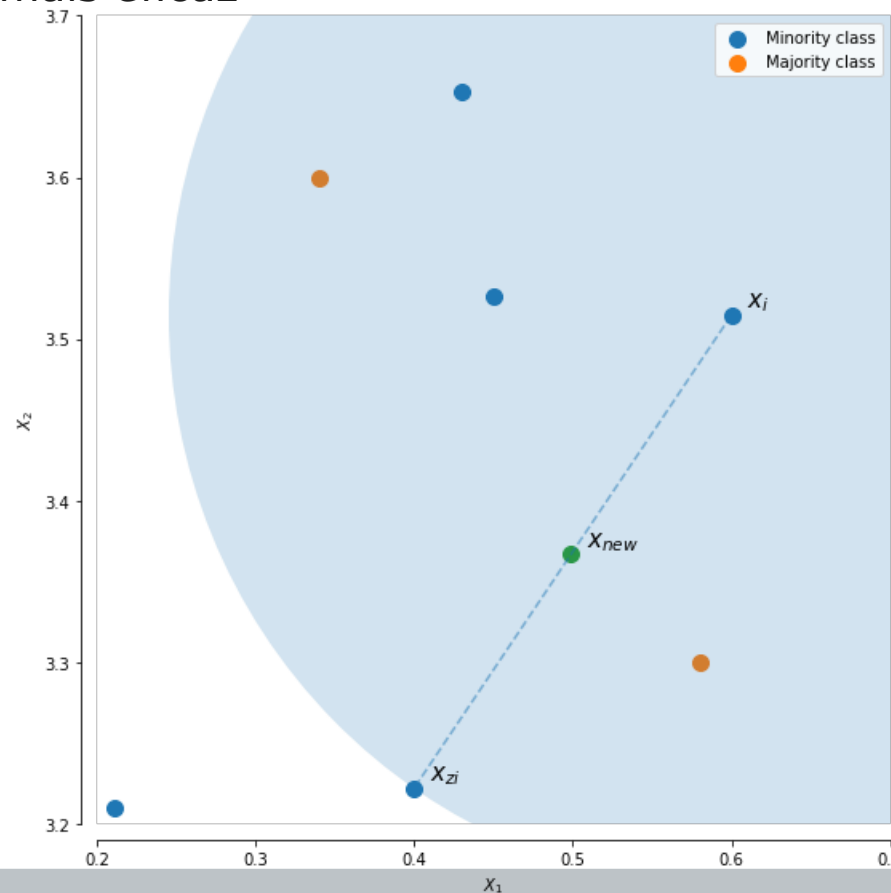
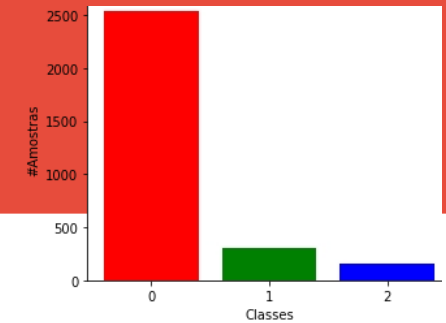
Under/Oversampling





# Dados Desbalanceados

- Replicação não é necessariamente eficaz
  - Não aumenta a representatividade
- Interpolação pode ser mais eficaz



# Dados Desbalanceados

- Balancear sinteticamente nem sempre é a solução:
  - Introdução de padrões artificiais
  - Risco de overfitting
  - Avaliar Correlação
  - Não aumenta a representatividade
- Solução
  - Coletar mais dados
  - Engenharia de Características
    - Implementar novas variáveis
  - Etc



# Let's Code

- Siga o link dos tutorias abaixo:

- Dados Desbalanceados:

[Tópico\\_02\\_Aprendizado\\_Supervisionado\\_Dados\\_Desbalanceados.ipynb](#)

- Tratando Dados Desbalanceados:

[Tópico\\_02\\_Aprendizado\\_Supervisionado\\_Tratando\\_Dados\\_Desbalanceados.ipynb](#)

# Considerações Finais

- Modelos são sensíveis a dados desbalanceados
  - Overfitting
- Balancear os dados é uma saída, porém
  - Undersampling: Pode gerar poucas amostras e baixa representação
  - Oversampling: Pode não aumentar a representatividade
- Análise crítica é essencial
  - Sensibilidade do modelo
  - Treino / Test

